אלגוריתם השכן הקרוב ביותר
K-Nearest Neighbors Algorithm

במסגרת קורס דאטה סיינטיסט של ג'ון ברייס

רועי פולניצר, PDS, CFV, F.IL.A.V.F.A., FRM
בעלים ואקטואר ראשי של פירמת הייעוץ וההדרכה – שווי פנימי
מייסד ויו"ר לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA)
מייסד ומנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA)

תל אביב, 14 בפברואר 2020

הגדרה

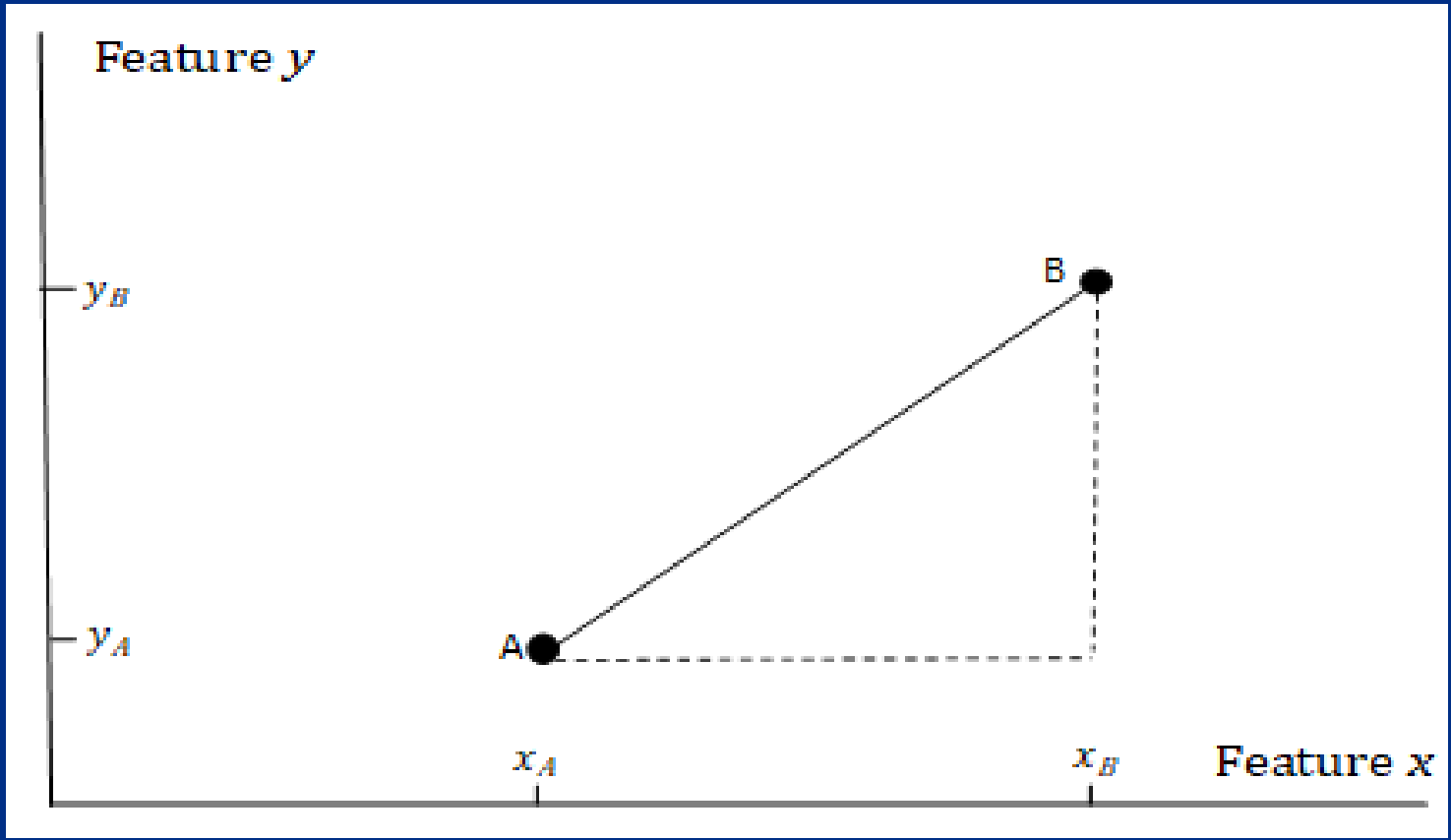
- אלגוריתם של למידת מכונה, שנועד לפתור בעיות חיזוי וגם בעיות סיווג.
- בעיות חיזוי, אלו הן בעיות שבהן אנו מנסים לחזות ערך רציף (כגון: מחיר, משקל, זמן וכו'). לדוגמא: תוצאת החיזוי יכולה להיות 5.38%.
- בעיות סיווג, אלו הן בעיות שבהן אנו מנסים לסווג תצפית מסוימת לאחת מכמה קטגוריות. לדוגמא: תוצאת הסיווג יכולה להיות סולבנטי או לא סולבנטי.

שלבי הפעולה

- שלב 1: בחירת ערך מסוים עבור k.
- שלב 2: מציאת k התצפיות במדגם שהמאפיינים שלהן הם הדומים ביותר למאפיינים של נשוא החיזוי/הסיווג – באמצעות נוסחת המרחק האוקלידי (קרי, "רמת הדימיון" נמדדת באמצעות מדד המרחק האוקלידי).
- שלב 3: ביצוע חיזוי/סיווג.

מדד המרחק האוקלידי (משפט פיתגורס)

$$Distance = \sqrt{(y_B - y_A)^2 + (x_B - x_A)^2}$$



דוגמא לבעיית חיזוי

- נניח שאנו רוצים לחזות את מחירה של דירה בשכונה מסוימת, על בסיס 2 מאפיינים בלבד:
  1. מספר חדרים;
  2. שטח במ"ר.

דוגמא לבעיית חיזוי - המשך

- נניח ובחרנו שרירותית k=3, או אז עלינו לקחת את המחירים של 3 הדירות הדומות ביותר מבחינת המאפיינים שלהן לאלו של הדירה הנחזית.
- נניח שידוע לנו שהמחירים של 3 הדירות הדומות ביותר הם: 2.30 מיליון ש"ח, 2.45 מיליון ש"ח ו- 2.18 מיליון ש"ח.
- לפיכך, מחירה של הדירה הנחזית, על בסיס אלגוריתם k-NN בעבור k=3, שווה ל- 2.31 מיליון ש"ח.

$$\frac{2.30 + 2.45 + 2.18}{3} = 2.31$$

דוגמא לבעיית חיזוי - המשך

- אילו היינו בוחרים k=2, הרי שהיה עלינו לקחת את המחירים של 2 הדירות הדומות ביותר מבחינת המאפיינים שלהן לאלו של הדירה הנחזית.
- נניח שידוע לנו שהמחירים של 2 הדירות הדומות ביותר הם: 2.30 מיליון ש"ח ו- 2.45 מיליון ש"ח.
- לפיכך, מחירה של הדירה הנחזית, על בסיס אלגוריתם k-NN בעבור k=2, שווה ל- 2.375 מיליון ש"ח.

$$\frac{2.30 + 2.45}{2} = 2.375$$

דוגמא לבעיית סיווג

- נניח שאנו רוצים להחליט האם לאשר בקשת הלוואה מסוימת, על בסיס 4 מאפיינים בלבד:
  1. בעלות על דירה (כן/לא);
  2. הכנסה שנתית;
  3. יחס חוב להכנסה;
  4. ציון אשראי.
- למעשה, אנו מתבקשים לסווג את ההלוואה לאחת מ- 2 קטגוריות: הלוואות טובות או הלוואה רעות.

דוגמא לבעיית סיווג - המשך

- נניח ובחרנו שרירותית k=10, או אז עלינו לקחת את התוצאות של 10 ההלוואות הדומות ביותר מבחינת המאפיינים שלהן לאלו של ההלוואה המסווגת.
- נניח שידוע לנו שמתוך אותן 10 ההלוואות הדומות ביותר: 8 התבררו כהלוואות טובות ו- 2 התבררו כהלוואות רעות.
- לפיכך, ההלוואה המסווגת, על פי אלגוריתם k-NN בעבור k=10, תשוייך לקטגוריית ההלוואות הטובות.

דוגמא לבעיית סיווג - המשך

- אילו היינו בוחרים k=2, הרי שהיה עלינו לקחת את 2 ההלוואות הדומות ביותר מבחינת המאפיינים שלהן לאלו של ההלוואה המסווגת.
- נניח שידוע לנו ש- 2 ההלוואות הדומות ביותר התבררו כהלוואות רעות.
- לפיכך, ההלוואה המסווגת, על פי אלגוריתם k-NN בעבור k=2, תשוייך לקטגוריית ההלוואות הרעות.

נשמח לעמוד לרשותכם

רועי פולניצר, PDS, CFV, F.IL.A.V.F.A., FRM

בעלים ואקטואר ראשי של פירמת הייעוץ וההדרכה – שווי פנימי

מייסד ויו"ר לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA)

מייסד ומנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA)

polanitz7@gmail.com

www.pdsia.org

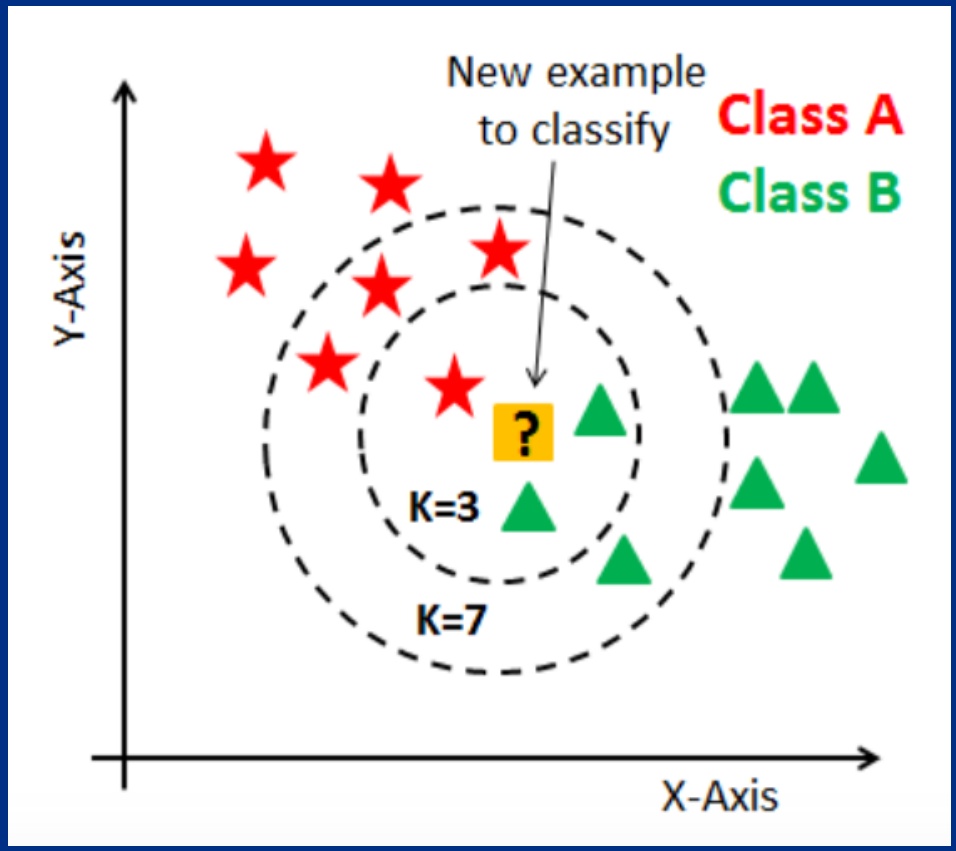
נספח 1 – נוסחת המרחק האוקלידי

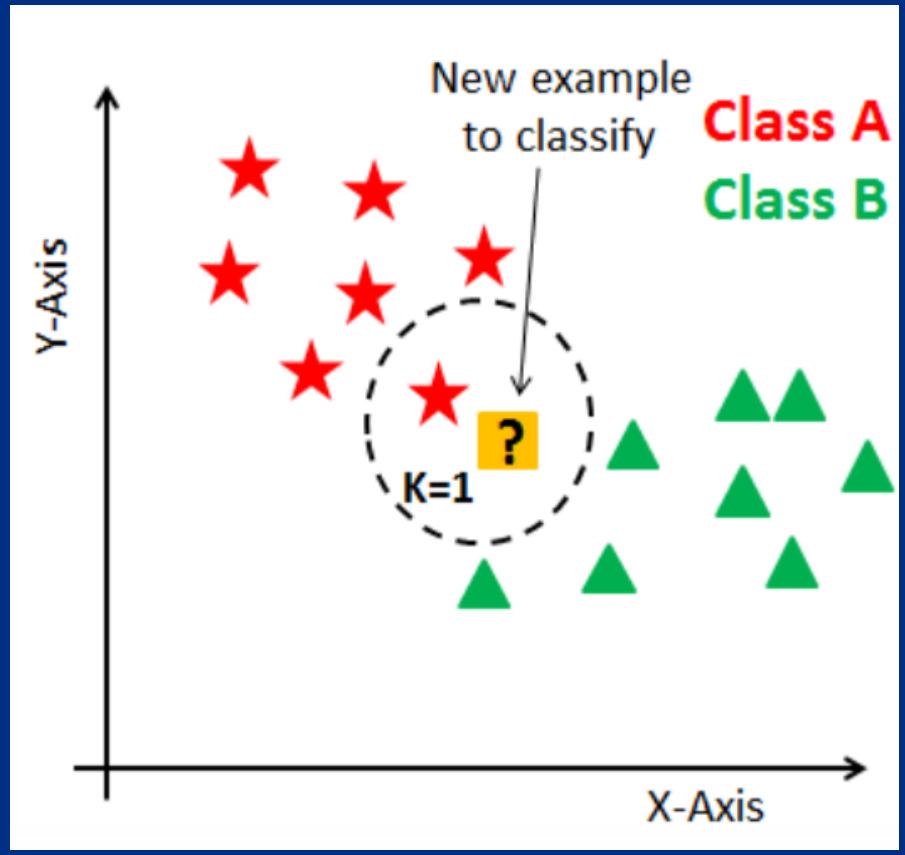
❖ ברגיל, כאשר ישנם m מאפיינים, המרחק האוקלידי בין נקודה P לנקודה Q הוא

$$\sqrt{\sum_{j=1}^{m}\left(v_{pj}-v_{qj}\right)^2}$$

כאשר v_{pj} ו- v_{qj} הם הערכים של המאפיין ה- j עבור P ו- Q, בהתאמה

נספח 2 – המחשה ויזואלית





נספח 3 – קריאה נוספת

פולניצר, ר' (2019), "חיזוי באמצעות שיטות K Nearest Neighbors", סטטוס - כתב עת לחשיבה ניהולית ואסטרטגית, דצמבר.